Analysis of host response to bacterial infection using error model based gene expression microarray experiments (Erratum in NAR 33 (7) 2352-3)

Stekel, Dov; Sarti, D; Trevino, V; Zhang, Lihong; Salmon, Michael; Buckley, Christopher; Stevens, M; Pallen, Mark; Penn, Charles; Falciani, Francesco

DOI:
10.1093/nar/gni050

License:
Other (please specify with Rights Statement)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Stekel, D, Sarti, D, Trevino, V, Zhang, L, Salmon, M, Buckley, C, Stevens, M, Pallen, M, Penn, C & Falciani, F 2005, 'Analysis of host response to bacterial infection using error model based gene expression microarray experiments (Erratum in NAR 33 (7) 2352-3)', *Nucleic Acids Research*, vol. 33, no. 6, e53. https://doi.org/10.1093/nar/gni050

Nucleic Acids Research, 2005, Vol. 33, No. 6 ***e53***
doi:10.1093/nar/gni050

Analysis of host response to bacterial infection using error model based gene expression microarray experiments

Dov J. Stekel, Donatella Sarti, Victor Trevino, Lihong Zhang[2], Mike Salmon[1], Chris D. Buckley[1], Mark Stevens[3], Mark J. Pallen[2], Charles Penn and Francesco Falciani*

School of Biosciences and [1]MRC Centre for Immune Regulation, Division of Immunity and Infection and [2]Bacterial Pathogenesis and Genomics Unit, Division of Immunity and Infection, Medical School, The University of Birmingham, Birmingham B15 2TT, UK and [3]Institute of Animal Health, Compton, UK

Received October 27, 2004; Revised December 8, 2004; Accepted March 1, 2005

ABSTRACT

A key step in the analysis of microarray data is the selection of genes that are differentially expressed. Ideally, such experiments should be properly replicated in order to infer both technical and biological variability, and the data should be subjected to rigorous hypothesis tests to identify the differentially expressed genes. However, in microarray experiments involving the analysis of very large numbers of biological samples, replication is not always practical. Therefore, there is a need for a method to select differentially expressed genes in a rational way from insufficiently replicated data. In this paper, we describe a simple method that uses bootstrapping to generate an error model from a replicated pilot study that can be used to identify differentially expressed genes in subsequent large-scale studies on the same platform, but in which there may be no replicated arrays. The method builds a stratified error model that includes array-to-array variability, feature-to-feature variability and the dependence of error on signal intensity. We apply this model to the characterization of the host response in a model of bacterial infection of human intestinal epithelial cells. We demonstrate the effectiveness of error model based microarray experiments and propose this as a general strategy for a microarray-based screening of large collections of biological samples.

INTRODUCTION

DNA microarrays are devices that measure the expression of many thousands of genes in parallel (1). They have revolutionized molecular biology, and in the last five years, their use has grown very rapidly throughout academia, medicine, and the pharmaceutical, biotechnology, agrochemical and food industries (2–4).

One of the common aims of microarray experiments is to identify genes that are differentially expressed in one set of tissues or cells relative to another. Typically, these may be diseased versus normal tissue (5,6), treated versus untreated cells (7,8) or wild-type versus mutant strains of organisms (9,10).

When such data are analysed, it is normal to apply methods to one gene at a time, and then to rank the genes according to some measure of the extent to which the gene is differentially expressed. In the earliest microarray experiments, researchers used the fold ratio to describe the level of differential gene expression. More recently, it has become increasingly common to use more rigorous statistical analyses, such as *t*-tests, bootstrap tests or ANOVAs (11,12).

The use of the hypothesis-testing framework of classical statistics is unquestionably the best method to analyse microarray data for differentially expressed genes. The reason for this is that these methods allow the scientist to make a statistical inference from the data: an extrapolation from the individuals being studied to the population from which the individuals derive. However in order to make this inference, we need to run an experiment with sufficiently many biological replicates (i.e. individuals from the population of study) so that the statistical analyses can provide reliable results.

*To whom correspondence should be addressed. Tel: +44 121 4143037; Fax: +44 121 4145925; Email: f.falciani@bham.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Many microarray-based projects now include the analysis of hundreds or thousands of biological samples (9). In this case, it is often impractical to perform experiments with a sufficient number of experimental replicates. In such experiments, we were not able to make a statistical inference from the data, and thus are only able to make assertions about the gene expressions measured in the sample in the study.

In the absence of proper replication we would, however, still like to apply some rational criteria to identify differentially expressed genes so that we can generate biological hypotheses from the data. For poorly replicated experiments, it is still possible to estimate variance from the available information. Recently Jain *et al*. (13) have developed the local pool error (LPE) test, a methodology that allows better estimates of the experimental variability from poorly replicated experiments. This method has been applied to a simple model of T-cell activation and has been shown to be effective in experiments based on three replicates. However, in very large expression profiling experiments it may be impractical to perform any replication at all. A suitable strategy in this case would be to construct an error model from an initial highly replicated experiment (e.g. testing a reference mutant strain) and apply this to the analysis of all other samples in a non-replicated design. A number of these approaches have been developed and applied to the analysis of microarray experiments. Both GeneSpring (14) and Rosetta (15) use this approach in their data analysis software. However, there are two disadvantages with their approaches. First, they base their approaches on normal distributions (16). Typically, the errors in microarray data are not normally distributed (17,18), and in this paper we show how the use of bootstrap distributions produces more realistic error models than normal distributions. Second, they generate *P*-values for each gene. This is problematic because the large number of genes on a microarray results in false positive results and, hence, the *P*-values are difficult to interpret.

The aim of this paper is to show how to generate an error model from a replicated pilot study, which can then be applied to subsequent experiments on the same platform to help select differentially expressed genes. As with other error model based methods, this method is applicable only when subsequent experiments are performed using a similar experimental system, and so are likely to have a similar error structure as the pilot experiment. Our model uses a bootstrap distribution, which allows for errors that depend on signal intensity, and includes array-to-array and feature-to-feature variability, as well as the dependence of error on encoded spot features (in this example on failed features).

The false discovery rate (FDR) (19) is of great importance in microarray experiments because of the large number of genes that are being tested. In traditional experiments, we may only be measuring one variable; when we see a positive result, such as differential gene expression, we would normally attribute this to genuine scientific effect. In a microarray experiment, we may be analysing many thousands of genes in parallel. Because of the large number of genes being analysed, there is an increased likelihood that some genes that are not differentially expressed will appear to be so, as a result of experimental errors in the measurements. Therefore, we include an FDR calculation and use this as the criterion for selecting differentially expressed genes.

We demonstrate the method by building an error model for a human array representing an unbiased selection of 850 genes involved in key biological processes. This model has been applied to the characterization of the host response in an established *in vitro* model of bacterial infection. We validate our method with respect to genes expected to be differentially regulated in response to infection [nuclear factor κB (NF-κB) pathway] and show that we can detect 65% of expected differences at an FDR threshold of 10%. This compares favourably with the most widely used error model based method that detects only 40% of these genes. We then describe the biological implications of the experimental results and demonstrate that important biological conclusions can be derived from the proposed analysis strategy.

SYSTEMS AND METHODS

The biological system

Escherichia coli typically colonize the gastrointestinal tract of the human intestine within a few hours of birth. This results usually in a mutually beneficial relationship that lasts for life. However, there are some strains of *E.coli* that have acquired a number of virulence factors that confer the ability to colonize new niches and result in a broad spectrum of diseases. This paper focuses on the analysis of the response of a human intestine epithelial cell line (Caco-2) to a number of enterohaemorrhagic *E.coli* (EHEC) and enteropathogenic *E.coli* (EPEC) strains.

Despite the relatively large differences in their gene complement (20), these two bacterial pathotypes share many of the virulence genes required for colonization. Some of these genes are clustered in chromosomal pathogenicity island termed the locus of enterocyte effacement (LEE). LEE encodes a type III protein secretion system, which serves to inject a set of bacterial proteins into host target cells (21). These bacterial expressed proteins induce a number of drastic transformations in the target cells, including cytoskeletal rearrangements [an important step in the formation of the typical attaching and effacing (AE)-lesion] (22,23), affected integrity of tight junctions and reduced transepithelial resistance *in vitro* (24) and induction of apoptosis (25). At the molecular level, both EHEC and EPEC strains are known to be potent inducers of NF-κB with consequent up-regulation of NF-κB downstream genes. This property of EHEC- and EPEC-infected cells is useful for the validation of our methodology since it provides a set of genes expected to be up-regulated during infection.

Despite the remarkable similarity in both the initial steps of colonization and the effect on host cell physiology, there are substantial differences between the EHEC and the EPEC pathotypes. An important difference is the production of Shiga-like toxins by the EHEC strain. These proteins are responsible for the bloody diarrhoea and haemorrhagic colitis induced by infection of EHEC (26), and act as powerful protein synthesis inhibitors (27). Less dramatic differences have also been demonstrated in other factors involved in colonization. Among these, the lymphocyte inhibitory factor (LifA) is a good example. LifA has been first discovered as a EPEC-secreted lymphotoxin capable of inhibiting the production of cytokines and proliferation in human peripheral blood mononuclear cells (PBMCs) (28,29). On the other hand, the EHEC

homologue of LifA (Efa1) has been identified independently as a protein involved in adhesion (30). It is still unclear if LifA alone has also a role in EPEC adhesion.

Experimental methods

Microarray design. An array of 850 core genes was constructed using 60mer synthetic oligonucleotides, which were then tested for hybridization efficacy and specificity. The array covers genes associated with apoptosis, cell cycle, senescence, chemokines, cytokines, receptors, adhesion and matrix and intracellular signalling pathways.

Immunofluorescence assay. Cells were seeded in 12-well plates and grown to confluence on glass cover slips, previously treated with nitric acid for 10 min and then washed with water and ethanol. After the infection, the cells were processed and stained using Phalloidin conjugated with Alexa 488 (Molecular Probe, The Netherlands) as described by Knutton *et al*. (31).

Infection procedure. This study employs the following bacterial strains: EPEC O127:H6 E2348/69, EPEC O127:H6 E2348/69 ΔlifA, EHEC O157:H7 Sakai ($stx^{-/-}$) and EHEC O157:H7 Sakai ($stx^{-/-}$) Δler. Individual bacterial colonies were grown overnight at 37°C without shaking in Luri–Bertani broth. The overnight culture was diluted 1:30 in DMEM (GIBCO, UK) and incubated at 37°C and 5% CO_2 to mid-log phase. Bacteria at a multiplicity of infection (MOI) of 50 were added to semi-confluent layers of Caco-2 cells (ATCC, USA) grown in DMEM supplemented with 10% FBS (GIBCO). After infection, unbound bacteria were washed with PBS and Caco-2 cell RNA was extracted using RNAasy (Qiagen, UK) according to the manufacturer's instructions.

Adhesion assay. Caco-2 cells were seeded in six-well plates, grown to confluence and infected as described above. After the infection, Caco-2 cells were washed with PBS, added with fresh media and the infection was continued for further 2.5 h. Then the host cells and the adherent bacteria were washed with PBS, fixed with methanol and stained with Giemsa solution. The percentage of cells infected was assessed from a minimum of 10 different microscopic fields.

Inactivation of bacteria. Bacteria were grown as described in the infection procedure, washed twice in PBS, resuspended in 1.5% formalin in PBS, and incubated for 1.5 h at 23°C. Subsequently, bacteria were washed in PBS and heat inactivated for 5 min at 80°C.

Microarray technology. An aliquot of 30 µg of total RNA was labelled by direct incorporation of dCTP, conjugated with Cy3 and Cy5. Probe synthesis, purification, and microarray hybridization and washing have been conducted as described previously (32). Slides were scanned using an Axon scanner (Axon, USA) and image analysis has been performed using the software Genepix v3.0 (Axon). Each individual slide has been hybridized to directly compare RNA from infected cells (typically labelled with Cy3) against RNA from uninfected cells grown in identical conditions (typically labelled with Cy5).

Computational methods

Normalization. We have not used features that have been flagged by the image processing software (GenePix). The data have been normalized by taking log to base 2 and computing the log ratio of infected to uninfected for each replicate feature on the array. Systematic dye bias was corrected by a loess fit of the log ratio against the average signal intensity, and the loess-fitted curve was subtracted from the log ratios to produce normalized log ratios (12,33,34). A measurement of log ratio was computed for each gene on each array by taking the average of the normalized log ratios of the replicate features. A final measurement of log ratio for each gene was calculated by taking the average of the array replicates. These have been ordered according to magnitude. As some measurements are missing (owing to bad features), the number of replicate features and replicate arrays for each measured gene expression ratio may vary.

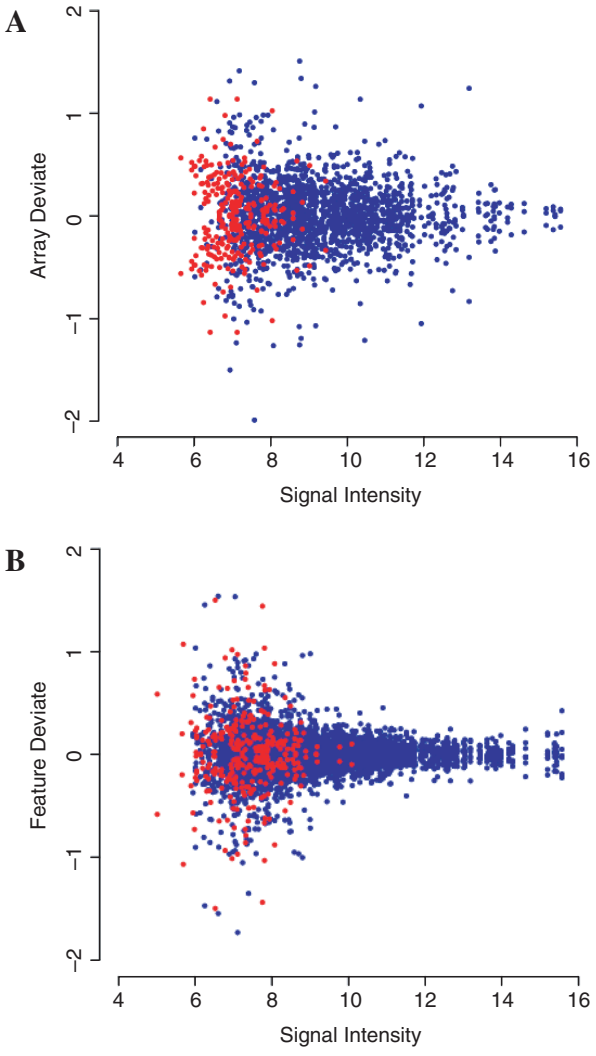
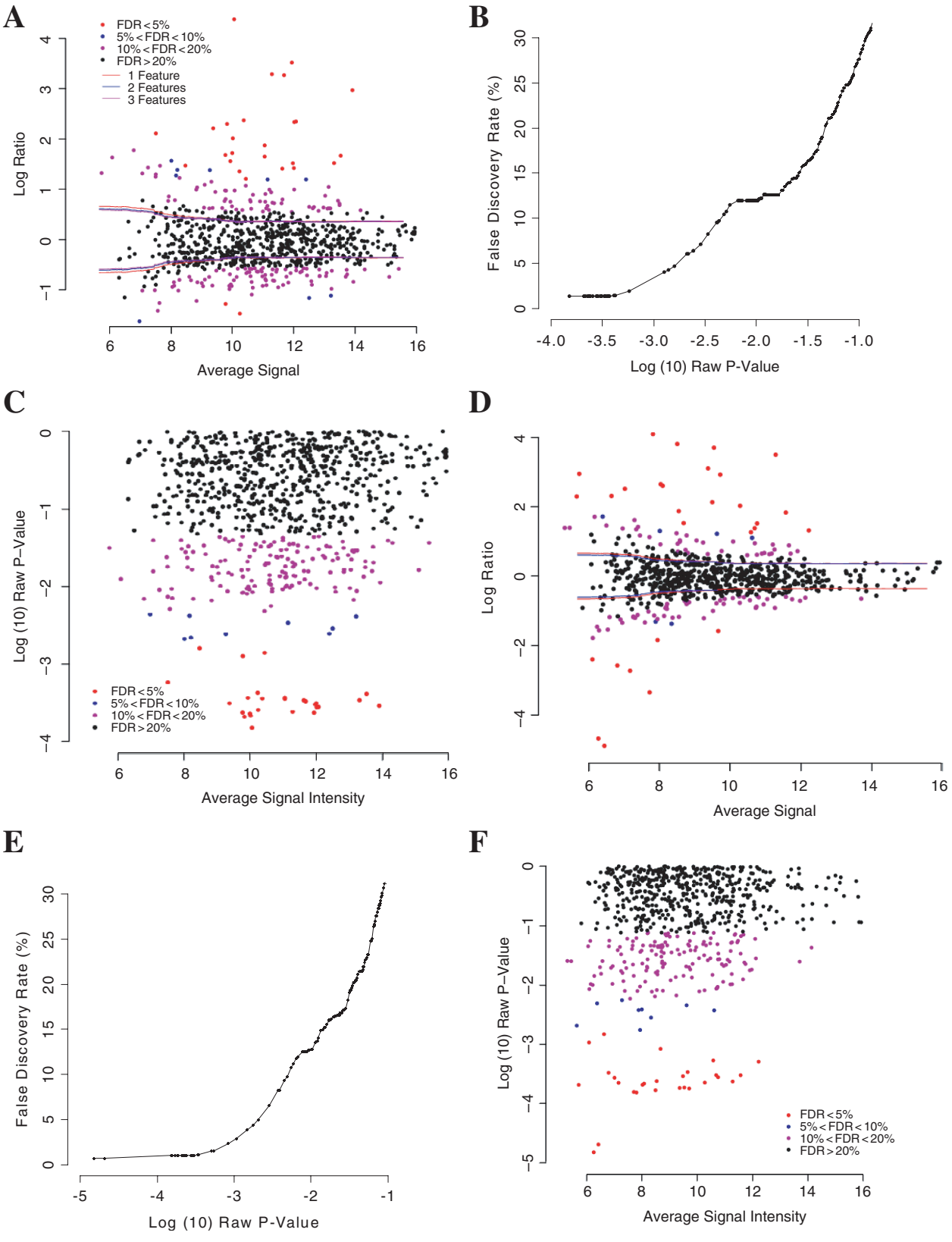


Figure 1. (**A**) Distribution of array deviates as a function of signal intensity. Red dots are the genes with only two successful arrays; blue dots are genes with three successful arrays. The magnitude of the errors depends on signal intensity, with larger errors at low-signal intensity, and smaller errors at high-signal intensity. Thus the distribution of errors is not log–normal, and so the error model requires an approach that includes dependence of error on signal intensity. The magnitude of the array deviates does not appear to depend on the number of successful arrays. (**B**) Similar plot for the feature deviates. The plot shows very similar behaviour, with a dependence of error on signal intensity. The magnitude of the feature deviates is slightly smaller than the array deviates.

Error model construction. In generating an error model, we use a bootstrap method that does not make any assumption about the structure of the variability, but instead uses the observed errors as the error distribution. Thus bootstrapping is appropriate for all error distributions, even when the errors are not normal. In this experiment, the situation is complex for three reasons. First of all, different genes have different numbers of successful replicates. Second, the magnitude of the



errors depends on the signal intensity. And third, there are two levels of error: array-to-array variability and feature-to-feature variability.

The bootstrap error model represents synthetic genes that are not differentially expressed, but which have log ratios different from zero because of array-to-array and feature-to-feature variabilities. Since we are interested in applying the error model to genes for which there may be one or two failed features, we build three error models: one for genes for which all three features have been successful, one for genes with two successful features and one for genes with just one successful feature.

We build the error models using the gene expression values of the genes in the pilot experiment. For each gene in the pilot study, we compute the array and feature deviates as follows: the array deviates are the differences between the average log ratio of the features for that gene on the array and the log ratio for that gene averaged across all three arrays; the feature deviates are the differences between the log ratios of each feature and the average log ratio of all corresponding features on the same array.

The error model is constructed by adding feature deviates to array deviates. Since there are many more feature deviates than array deviates, we construct a distribution that is a hybrid between a permutation distribution and a bootstrap distribution (35). As the errors depend on signal intensity, we construct a different bootstrap distribution for each intensity level—corresponding to each gene in the pilot experiment. In order to do this, we need to construct a set of bootstrap errors associated with each possible intensity level. For each gene, there are either two or three array deviates, depending on the number of successful arrays. For each array deviate, we choose feature deviates at random from genes with a similar intensity level, by using a window of width 200 genes (after ordering the genes by intensity level). For the bootstrap distribution for arrays with one successful feature, we add one feature deviate to the array deviate; for the distribution with two successful features we add the average of two feature deviates to the array deviate; and for the bootstrap distribution for arrays with three successful features we add the average of three feature deviates to the array deviate. This generates a single value of the bootstrap distribution associated with that intensity level. We repeat this process 500 times for each array deviate, thus generating either 1000 or 1500 bootstrap ratios for each signal intensity level. These are used for generating the bootstrap distributions.

Identification of differentially expressed genes. Given a new microarray experiment on the same platform, we can use the error distribution to identify differentially expressed genes. For each gene, we look at the distribution of bootstrap log ratios for signal intensities within a window of 200 bootstrap genes with signal intensities similar to the gene being analysed, using the bootstrap distribution for the same number of features as the number of successful features for that gene (1, 2 or 3). Thus we have a distribution of up to 300 000 bootstrap log ratios, which represent errors at signal intensity similar to the gene in question. We compare the log ratio of the gene with the 300 000 bootstrap log ratios, and count the number of bootstrap log ratios that are more extreme (either positive or negative) than the gene being analysed. The bootstrap *P*-value is then the number of more extreme bootstrap log ratios divided by the size of the bootstrap distribution. In practice, we add 1 to the number of bootstrap log ratios to avoid getting *P*-values of 0. We select a threshold for differentially expressed genes using the FDR procedure (19).

The use of the FDR procedure also determines the number of bootstrap replicates to be used. In order to obtain any positive results, we need to ensure that the highest-ranked genes in the list (with the smallest *P*-values) can be significant. The smallest *P*-value obtainable is the reciprocal of the number of bootstrap replicates. Since the FDR is computed by multiplying the *P*-value by the number of genes on the array, we require the unadjusted *P*-value to be smaller than the desired FDR divided by the number of genes on the array. In this experiment, there are ~1000 genes on the array. Thus to allow possible FDRs of 1%, we require an analysis that can give an unadjusted *P*-value at least as small as 10^{-5}. In order

Figure 2. (**A**) MVA plot for the array for infection with the EPEC strain after 6 h. Each gene is represented by one spot, which is colour coded according to the FDR associated with its *P*-value. The fold ratio at which genes are called differentially expressed depends on its signal intensity—with genes at higher signal intensity being differentially expressed at lower log ratios than genes at low-signal intensity. This is a reflection of the intensity-dependent error model. On the same axes we have plotted the standard deviation of the intensity-dependent error model distributions. There are three distributions: one for genes with one successful feature; one for genes with two successful features and one for genes with all three successful features. The distributions are most different at low-signal intensities, where the feature deviates are similar in magnitude to the array deviates. This is also the range of intensity where features are likely to fail. At high-signal intensities, where the magnitude of the feature deviates is much less than the magnitude of the array deviates, the distributions are dominated by the array deviates, and are very similar. In reality, features are much less likely to fail at high-intensity levels, so there is only a real need for the distribution for three successful replicates. (**B**) FDR plot for the EPEC strain array. On the *x*-axis we plot the *P*-value of the genes on a log (base 10) scale; on the *y*-axis we plot the FDR associated with that gene. This is essentially the expected number of false positives (equal to the number of genes in the analysis multiplied by the *P*-value), divided by the observed number of genes with *P*-values less than or equal to the *P*-value (i.e. the rank of the gene with this *P*-value). On this array, there are 804 genes in the analysis. We use the FDR curve to select differentially expressed genes. There are 27 genes with FDR <5% and 36 genes with FDR <10%. (**C**) Plot of average signal intensity against *P*-value for the genes in the EPEC ΔlifA mutant array. This is a diagnostic plot to determine the performance of our error model. Each spot represents a gene, and has been colour-coded according to the FDR associated with its *P*-value. There is no dependence of *P*-value on signal intensity, suggesting that our error model is performing well with these data. Furthermore, the FDR thresholds also do not depend on signal intensity, again supporting the use of our error model with these data. This contrasts with the MVA plot of log ratio against signal intensity, where there are more extreme log ratios at lower signal intensities than at higher signal intensities. The use of fold-ratio thresholds, or any other approach that does not include dependence of error on signal intensity, would be inappropriate with these data. (**D**) MVA plot of log ratio against signal intensity for the array for the EPEC ΔlifA mutant after 6 h. The results on this array show a far greater dependence of log ratio on signal intensity, with many more extreme values at low intensity, and fewer extreme values at high intensity. As with the ΔlifA, the analysis selects differentially expressed genes at high-signal intensities with lower fold ratios than the differentially expressed genes at low-signal intensities. (**E**) FDR plot for the ΔlifA mutant array. The FDR shows a similar behaviour. The top 29 genes have FDR <5% and the top 35 genes have FDR <10%. (**F**) Diagnostic plot of *P*-value against signal intensity for the ΔlifA mutant array. In general, there is no dependence of *P*-value on signal intensity. Similarly, there is no dependence of the FDR on signal intensity. However, there are two genes (BCL-2 antagonist of cell death and RAR-e) in the bottom-left-hand corner of the plot with very low-signal intensity and *P*-values. From this plot, we would suspect that these genes are outliers and do not represent truly differentially expressed genes. Furthermore, both these genes have only one successful feature, indicating that these data are likely to be less reliable.

to ensure this, we have chosen to use bootstrap distributions of size 300 000.

Implementation of the error model. The methodology described above has been implemented both in the programming language Perl and in the statistical programming environment R (www.r-project.org). Libraries are available from the authors upon request.

Cluster analysis. Cluster analysis has been performed using hierarchical clustering with the unweighted pair-group method using the arithmetic averages (UPGMAs) on a similarity matrix built with Pearson correlation coefficient. Figure 1 has been produced using the Gepas tool set (36).

RESULTS

Error model

We have generated error models representative of a human microarray and applied this to study the host cell response to bacterial infection. The models we have developed are based on a three independent biological and technical replications of an infection of a colon tumor cell line Caco-2 with the EPEC 0127:H6 strain. Genes in the human microarray have been spotted three times in different areas of the slide to monitor within slide variability. This array type has been manufactured utilizing a widely used commercial oligonucleotide set developed and commercialized by Operon (Qiagen Ltd).

The error model depends on both the array-to-array and the feature-to-feature deviation (Figure 1). Both deviations are intensity-dependent, with greater variability at low-signal intensities than at high-signal intensities. There is slightly greater array-to-array variability (SD = 0.27) than feature-to-feature variability (SD = 0.17).

Since the deviations are intensity-dependent, the distribution of the error model itself is intensity-dependent. We plot the standard deviation for the error model as a function of intensity (Figure 2A and D). The model for one successful replicate has an SD ~0.66 for lowest expressed genes (average intensity of ~6); the SD decreases to ~0.36 for genes of moderate expression (average intensity of ~10), and remains approximately constant at this lower level for higher expressed genes. The models for genes with two and three successful replicates have very similar behaviour but with standard deviations starting at 0.61 and 0.59 for the lowest expressed genes, and then decreasing to similar values for the high-expressed genes.

Statistical verification

We verify our results by plotting the *P*-value as a function of signal intensity (Figure 2C and F). As the error model includes a dependence of error on signal intensity, there should be no bias for differentially expressed genes at any level of signal intensity. In practice, on some arrays this is the case (Figure 2C), but on arrays where there is a more pronounced dependence of error on signal intensity (Figure 2D), some of the differentially expressed genes appear suspicious on this plot (Figure 2F) and are likely to be outliers.

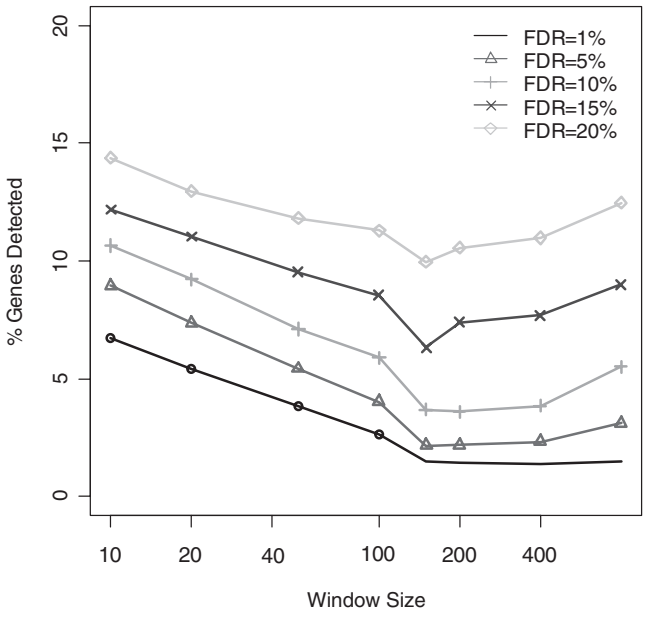


Figure 3. Choice of the window width. The figure shows the percentage of genes detected (ordinate) as significant at given FDR as a function of window width (abscissa). It can be noted that the percentage of genes detected stabilizes around $D = 150$ independently of the FDR threshold.

Window width

We have investigated the influence of the window width on the number of genes detected as differentially expressed for a given FDR threshold. Figure 3 displays the percentage of genes detected as differentially expressed (ordinate) as a function of window width (abscissa). Our analysis shows that the number of genes detected with a window width <150 is unstable and tend to decrease with the increase in the window width. The number of genes is then stable up to a window size of 400 for a range of FDR values.

Characterization of EPEC and EHEC infection

In order to evaluate the potential of error model based expression analysis, we have characterized the response of human Caco-2 intestinal epithelial cells to *E.coli* infection.

As stated previously four different bacterial strains were analysed using microarray technology. These were EHEC O157:H7 ($stx^{-/-}$), EPEC O127:H6, EHEC O157:H7 ($stx^{-/-}$) Δler and EPEC O127:H6 ΔlifA. In addition, the O157:H7 ($stx^{-/-}$) and EPEC O127:H6 strains were inactivated to test the ability of dead bacteria to elicit a host response.

In order to verify that a comparable number of human cells would be infected in our cultures, we have performed an adhesion assay and estimated the percentage of cells infected by each strain of *E.coli*. The assay has shown that in our experimental conditions a very similar number of Caco-2 cells are infected with the wild-type strains of EHEC and EPEC cells. In order to verify the success of the infection, we have also performed a fluorescent staining of the actin cytoskeleton of infected cells (Figure 4). Our analysis, in accordance with the recent findings by Cleary *et al.* (37), reveals dramatic difference between actin rearrangements of EHEC- and EPEC-infected cells. EHEC cells rearrange the actin

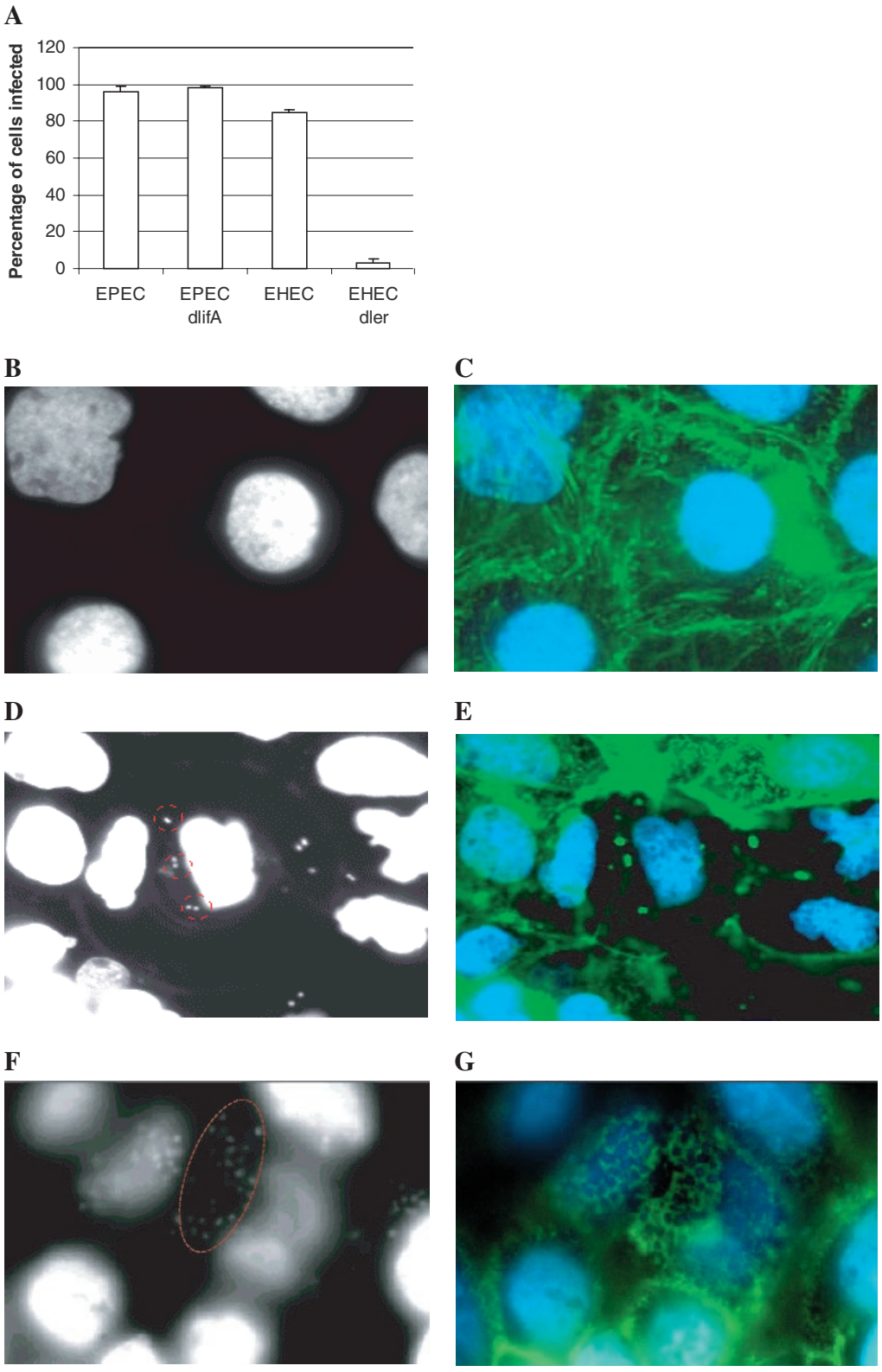
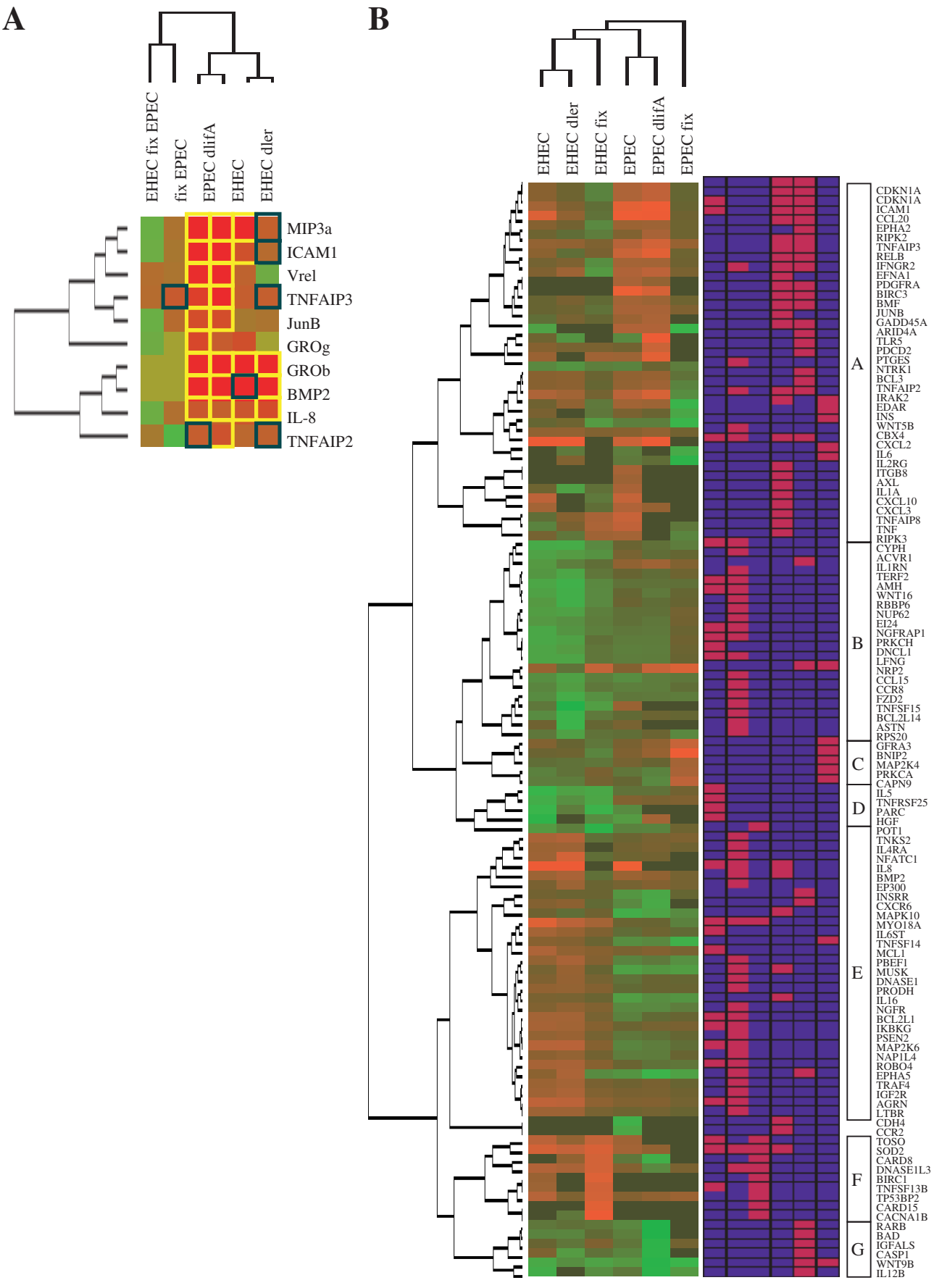


Figure 4. Characterization of EPEC and EHEC infection. (**A**) The percentage of Caco-2 cells infected with EPEC O127:H6 (EPEC), EHEC O157:H7 Sakai ($stx^{-/-}$) (EHEC), EPEC O127:H6 ΔlifA (EPEC dlifA) and EHEC O157:H7 Sakai ($stx^{-/-}$) Δler (EHEC dler) is shown. The graph clearly shows that the majority of cells are infected in all strains tested except the EHEC Δler. (**B**–**G**) The result of immunofluorescence assay is shown. The images of Caco-2 cell infected with EPEC O127:H6 and EHEC O157:H7 Sakai ($stx^{-/-}$) for 2 h are represented respectively in (F) and (G), and (D) and (E). Control Caco-2 cells (non-infected) are shown in (B and C). In (B, D and F), only the DAPI fluorescence is shown, whereas in (C, E and G) the merged fluorescence of both phalloidin (staining the cytoskeleton) and DAPI is shown. Red dashed circles in (D and F) indicate the position of the bacteria. In control cells actin microfilaments are diffused through the entire cell, while in infected cells they are clustered underneath the bacteria. It is noticeable that the morphology of the remodelling induced by EPEC O127:H6 and EHEC O157:H7 Sakai ($stx^{-/-}$) was different.


A
EHEC fix EPEC
fix EPEC
EPEC dlifA
EHEC
EHEC dler
MIP3a
ICAM1
Vrel
TNFAIP3
JunB
GROg
GROb
BMP2
IL-8
TNFAIP2
B
EHEC
EHEC dler
EHEC fix
EPEC
EPEC dlifA
EPEC fix
A
B
C
D
E
F
G
CDKN1A
CDKN1A
ICAM1
CCL20
EPHA2
RIPK2
TNFAIP3
RELB
IFNGR2
EFNA1
PDGFRA
BIRC3
BMF
JUNB
GADD45A
ARID4A
TLR5
PDCD2
PTGES
NTRK1
BCL3
TNFAIP2
IRAK2
EDAR
INS
WNT5B
CBX4
CXCL2
IL6
IL2RG
ITGB8
AXL
IL1A
CXCL10
CXCL3
TNFAIP8
TNF
RIPK3
CYPH
ACVR1
IL1RN
TERF2
AMH
WNT16
RBBP6
NUP62
EI24
NGFRAP1
PRKCH
DNCL1
LFNG
NRP2
CCL15
CCR8
FZD2
TNFSF15
BCL2L14
ASTN
RPS20
GFRA3
BNIP2
MAP2K4
PRKCA
CAPN9
IL5
TNFRSF25
PARC
HGF
POT1
TNKS2
IL4RA
NFATC1
IL8
BMP2
EP300
INSRR
CXCR6
MAPK10
MYO18A
IL6ST
TNFSF14
MCL1
PBEF1
MUSK
DNASE1
PRODH
IL16
NGFR
BCL2L1
IKBKG
PSEN2
MAP2K6
NAP1L4
ROBO4
EPHA5
TRAF4
IGF2R
AGRN
LTBR
CDH4
CCR2
TOSO
SOD2
CARD8
DNASE1L3
BIRC1
TNFSF13B
TP53BP2
CARD15
CACNA1B
RARB
BAD
IGFALS
CASP1
WNT9B
IL12B

cytoskeleton with evident polymerization below the bacterial cells, whereas in EPEC-infected cells the actin cytoskeleton is rearranged forming rings surrounding the bacterial cells. These differences in the phenotype of host infected cells could be just one aspect of more profound molecular differences in the host response.

The analysis of NF-κB downstream genes shows the efficacy of error model based expression profiling

The primary objective of our analysis has been to identify genes differentially expressed between control and infected cells. In particular, we first wanted to determine if error model based expression analysis is efficient in identifying genes expected to be differentially regulated consequent to infection. In order to do this we have selected, from our array, genes downstream the NF-κB pathway (38–40) (therefore known to be differentially regulated during infection) and we have chosen two arbitrary thresholds of FDR $\leqslant$10% and FDR $\leqslant$20% to generate lists of up- or down-regulated genes (Figure 5A). At a threshold of FDR $\leqslant$10%, significant up-regulation of NF-κB downstream genes has been detected in 65% of the expected cases (13 out of 20 cases; corresponding to the number of NF-κB assayed genes in the EHEC- and EPEC-infected cells). At an FDR $\leqslant$20%, the percentage of significant positive ratios slightly increases to 70%. In our experimental system, inactivated bacteria seem to be unable to activate the NF-κB pathway (Figure 5B). These results suggest that our methodology is accurate in identifying truly differentially expressed genes.

Global analysis of host response

In order to analyse the global response of cells in response to infection and to assess if error model analysis could provide new information on this biological process, we have then chosen an arbitrary threshold of FDR $\leqslant$10% to generate lists of genes significantly up- or down-regulated. We have then selected genes that were differentially regulated in at least one of the six arrays. This has led to a list of 116 genes. In order to facilitate the interpretation of our results we have performed cluster analysis on the 116 genes (Figure 5B). In addition to the conventional dendrogram and heat map representations, we have also produced a heat map labelling gene associated with an FDR value above the chosen threshold.

Interestingly, our analysis reveals that the transcriptional response of host cells infected by the EHEC strains is extremely divergent with respect to the response of cells infected with the EPEC strains. Most of the clusters that are identified represent strain-specific transcriptional responses as highlighted by the FDR map in Figure 5B. The three major clusters are labelled in Figure 5B as Cluster A, Cluster B and Cluster E. Cluster A represents genes that are up-regulated specifically with the infection of the EPEC strains, Cluster B represents genes down-regulated in response to EHEC infection and Cluster E genes are those that are primarily up-regulated in response to infection with the EHEC strains. Smaller groups represent genes that are activated in response to exposure to inactivated bacteria (Cluster C and Cluster F). Within the major groupings there are some interesting patterns. Although the response to infection with the EHEC Δler mutant strain was qualitatively similar to the response to the infection with the wild-type strain, it involved a much larger number of genes. This suggests that Ler or genes downstream of Ler may be responsible for silencing components of the host response. Host cell infection with the EPEC ΔlifA on the other hand produced an almost overlapping effect with respect to infection with the EPEC wild type, highlighting the limited role of LifA in controlling the response of epithelial cells.

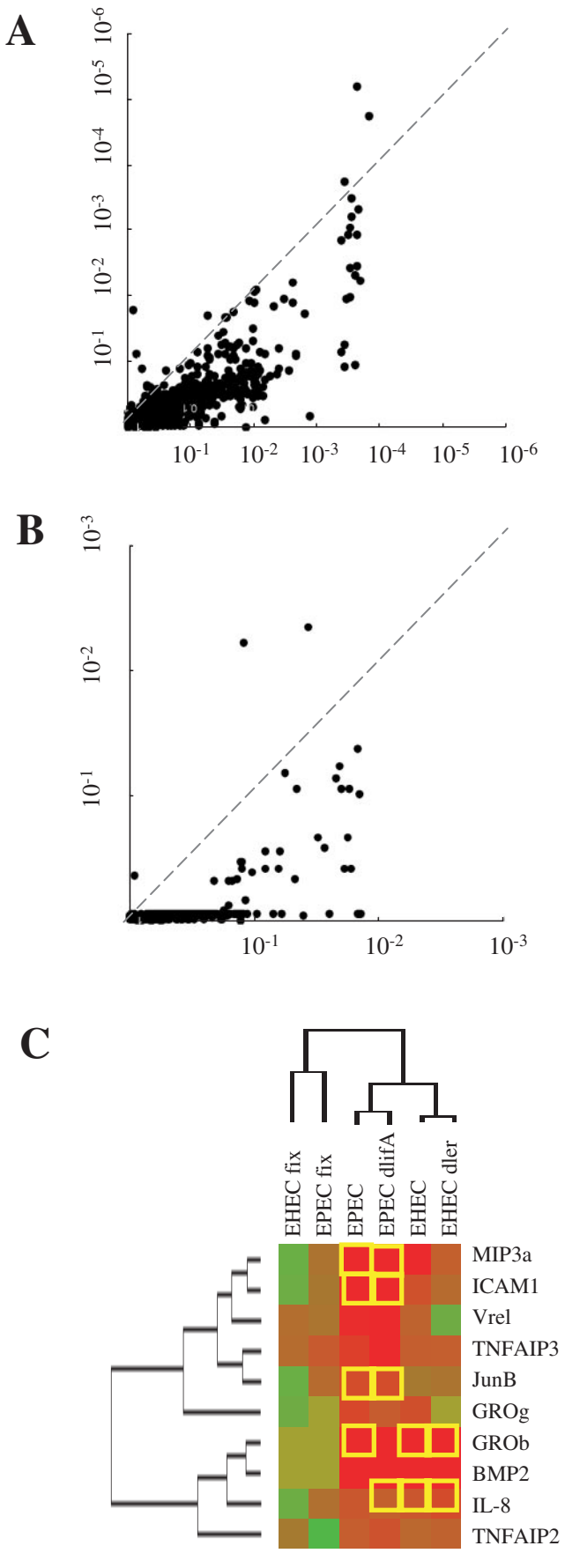
Genes that are specifically up-regulated in response to the EPEC infection include factors that are involved in cell cycle regulation (CDKN1A, RBP1, GADD45 and JUNB), apoptosis (JUNB, BIRC3 and BMP2) and two membrane receptors (IFNGR2 and PDGFRA). Interestingly, genes down-regulated in response to EHEC infections are characterized by a pro-apoptotic function (*NGFRAP1*, *CYPH* and *PRKCH*) whereas three anti-apoptotic genes (*TOSO*, *MYO18A* and *MCL1*) are up-regulated. *AGRIN* and its receptor *MUSK* are also up-regulated by EHEC strains. MUSK activation by AGRIN results in the onset of a tyrosine-kinase cascade that leads to cytoskeletal rearrangements in muscle cells at the neuromuscular junctions (NMJs) (41) and in T lymphocytes (42). The concomitant activation of AGRIN and its receptor suggest an involvement of this pathway in AE lesions formation in cells infected by EHEC. The up-regulation of these factors seems, however, to be insufficient to induce any actin rearrangement in the absence of Ler downstream genes.

We have also discovered that four components of the Wnt signalling pathway are differentially regulated in response to infection either by EHEC or EPEC strains. These are the ligands Wnt-15 (down-regulated in cells infected with EPEC O127:H6 ΔlifA and EPEC O127:H6 fixed), Wnt-5b1 (down-regulated in cells infected with EPEC O127:H6 fixed) and Wnt-16 2 [down-regulated in cells infected with EHEC O157:H7 ($stx^{-/-}$) and EHEC O157:H7 ($stx^{-/-}$) Δler] and the receptor Frizzled-2 [down-regulated in cells infected with EHEC O157:H7 ($stx^{-/-}$) Δler]. Since different members of the *Wnt* gene family are regulated by the different strains, it is possible that down-regulation of this morphogenetic pathway may be part of a general mechanism of host response. The response of host cells to exposure to inactivated bacteria seems to be almost exclusively associated with the induction of pro-apoptotic genes.

Comparison with the Rocke–Lorenzato two-component error model

Our methodology allows the identification of differentially expressed genes from unreplicated microarray experiments.

Figure 5. Cluster analysis. (A and B) The results of a two-way hierarchical clustering are shown. Samples are: Caco-2 cells infected for 6 h with EPEC O127:H6 (EPEC), EPEC O127:H6 ΔlifA (EPEC dlifA), EHEC O157:H7 Sakai ($stx^{-/-}$) (EHEC), EHEC O157:H7 Sakai ($stx^{-/-}$) Δler (EHEC dler) and EHEC O157:H7 Sakai ($stx^{-/-}$) (EHEC fix) and EPEC O127:H6 (EPEC fix) fixed; versus uninfected control Caco-2 cells. (**A**) The results of clustering using a subset of genes known to be downstream to NF-κB activation are shown. Highly significant ratios (FDR $\leqslant$10%) are marked in the heat map by yellow boxes. Significant genes (FDR >10% and <20%) are marked by black boxes. (**B**) The results of a two-way hierarchical clustering of genes that is differentially expressed (FDR $\leqslant$10%) in at least one of the arrays are shown. Dendrograms and heat maps are flanked by a colour-coded map representing genes associated with an FDR above (blue) and below (red) the chosen threshold in each array.



There are currently other two methods available; of these, the Rocke–Lorenzato two-component error model (16) is widely used in the microarray community because of its availability within the popular software application Genespring. In order to further validate our approach, we have performed a comparison between our methodology and the Rocke–Lorenzato model.

The Genespring implementation of the Rocke–Lorenzato model infers the experimental error by assuming that the level of variability is a function of the control signal strength within all the measurements in the array (16). In our analysis, a Rocke–Lorenzato has been fitted with data from each individual array and the FDR calculated as described by Benjamini and Hochberg (19).

With the exception of two genes, our methodology consistently estimates lower *P*-values respect to the Rocke–Lorenzato model (Figure 6A). This property is propagated to the FDR and is reflected in the number of genes the two methods detect as differentially regulated. For example, at a FDR threshold of 10%, the Rocke–Lorenzato model gives fewer genes differentially expressed (4 genes differentially expressed in the EHEC infected cells and 10 genes differentially expressed in the EPEC-infected cells) than our method (26 differentially expressed genes in the EHEC-infected cells and 33 differentially expressed genes in the EPEC-infected cells).

Moreover, the comparison of genes downstream NF-κB detected as differentially regulated by both methods reveal that our methodology is more effective in identifying genes known to be differentially regulated: 13 out of 20 NF-κB downstream genes are detected as differentially expressed with our methodology (Figure 5A) whereas only 8 out of 20 are detected as differentially regulated with the Rocke–Lorenzato model (Figure 6B).

DISCUSSION AND CONCLUSIONS

We have described a simple method that allows the selection of differentially expressed genes from microarray experiments with no replicates. This method relies on the prior construction of an error model from a replicated pilot experiment on the same platform, which is then applied to subsequent data that are produced. Although, in an ideal world, microarray experiments would have sufficient biological replicates to allow rigorous statistical analysis via hypothesis tests, we believe that this exercise is of great value, particularly in large-scale experiments where it is not financially viable to allow such replication. Of course, for this method to be applicable, the subsequent data must come from the experimental system similar to the pilot experiment.

The error model we derive uses a bootstrap distribution that is able to capture intensity-dependent variability,

Figure 6. Comparison with the Rocke–Lorenzato two-component error model. (**A**) Scatterplot comparing the *P*-values obtained with the Bootstrap error model (abscissa) with the *P*-values obtained with the Rocke–Lorenzato error model (ordinate) is displayed. (**B**) Scatterplot comparing the FDR obtained with the Bootstrap error model (abscissa) with the FDR obtained with the Rocke–Lorenzato error model (ordinate) is displayed. (**C**) The result of clustering using a subset of genes known to be downstream to NF-κB activation is shown. Ratios that are highly significant according to the Rocke–Lorenzato model (FDR ⩽10%) are marked in the heat map by yellow boxes. The map is directly compared with Figure 5A, which shows the results of the analysis on NF-κB downstream genes with the Bootstrap error model.

array-to-array variability, feature-to-feature variability, failed replicates and non-normal distributions of errors. Because of this, and its applicability to unreplicated experiments, we think that our model is superior to similar published works based on normal distributions. In particular, we demonstrated that our method performs substantially better than the widely used Rocke–Lorenzato model (16) implemented in the popular software package Genespring when applied to our data. Our work is also similar to the Locally Pooled Error model (13); however, our method has the advantage of being applicable to subsequent data where there is no replication at all.

We have analysed the effect of the window width in our error model. We have demonstrated that the number of genes detected at a given FDR threshold is stable in the range of 150–300 genes. The bootstrap model used in this analysis uses a fixed-width window of 200 in order to determine bootstrap distributions that capture the dependence of error on signal intensity. This value of window width is within the stable interval and it is sufficiently small to ensure a good inference of the intensity-dependent distribution.

We have based our gene expression threshold on an FDR calculation, so that we have been able to select differentially expressed genes with a quantitative measure of the likely number of false positives in the final list. An important issue in relationship with the detection of differentially expressed genes and the computation of FDRs is that there is substantial correlation structure in the gene expression profiles. There are modified forms of the FDR that attempt to take into account correlations in the data (43); however, the same group has found that the original FDR formulation is more effective for analysing microarray data (44). Therefore, we have used the original FDR formulation. An important area of future research into microarray data analysis will be to gain a greater understanding of the effects of the correlation structure on the selection of differentially expressed genes.

We have validated our method against a panel of genes known to be differentially regulated in Caco-2 cells and discovered that, at acceptable values of the FDR, up to 70% of the genes expected to be differentially expressed (after infection with alive bacteria) are detected as significant by our method.

Our observations that inactivated bacteria are unable to elicit NF-κB response are also consistent with reports from other groups (45,46). La Ferla *et al*. (45) have analysed a collection of 15 *E.coli* isolates and have discovered that the large majorities were able to activate the NF-κB pathway but none of them retained this ability after inactivation. Other groups have reported that inactivated *E.coli* cells are able to induce the expression of a number of NF-κB downstream genes in human PBMCs. Whether additional factors are required for the activation of this signalling pathway in our infection system is still unclear.

Our analysis has discovered a very divergent transcriptional response of the host cells in response to infection with EHEC or EPEC strains. These molecular differences reflect the divergent actin polymerization patterns that we have observed in EPEC- and EHEC-infected cells and are certainly much larger than originally expected. Moreover, the infection with an EHEC strain, mutated in the Ler regulator, seems to be associated with a larger transcriptional response with respect to cells infected with the wild-type strain, suggesting a broader role for Ler downstream genes in controlling host response than originally anticipated by Hauf and Chakraborty (47).

Our results can be used to generate new hypothesis on molecular pathways involved in the infection process. The observation that the interferon γ (IFN γ) receptor activation is a EPEC-specific component of the host response may be linked to the destruction of the IFN γ pathway in cells infected by EHEC (and not by EPEC) (48).

Our finding that the activation of the PDGF receptor is a specific event consequent to infection with EPEC, leads to another interesting hypothesis. PDGF stimulation leads to the activation of Abl tyrosine kinases that are known to be involved in the phosphorylation of the translocated protein Tir. The Tir protein is conserved in both EHEC and EPEC but the phosphorylation is required only in EPEC (49). The up-regulation of the PDGF receptor could create a favourable intracellular environment in cells infected by EPEC.

We have also observed two important pathways involved in cell remodelling regulated during infection. Agrin and its receptor is an example of a EHEC-specific pathway whereas the Wnt pathway may be a more general mechanism to control cellular remodelling during infection. The modulation of the Wnt signalling pathway may not have consequences only on the cell motility. The *WNT-16* gene is expressed in peripheral lymphoid organs and has been shown to be involved in haematopoiesis and to stimulate the proliferation of pre-B cells (50). The down-regulation of this gene by EHEC could therefore have an impact in the regulation of the host immune response at the mucosal level.

Error model gene expression analysis is not a substitute for experimental replication and proper statistical analysis. Indeed, because of the lack of biological replication, any hypotheses derived from these results would require further verification. However, our methodology is designed to be used in large-scale high-throughput screenings, where it would not be economically viable to generate the necessary level of replication. We have shown that our methodology can provide information on cell responses that is sufficiently detailed to allow hypothesis formulation, and thus that our method is amenable to be applied to the analysis of very large collections of biological samples. Therefore, this methodology makes it possible to perform meaningful analyses of microarray data from large-scale screenings of bacterial mutant collections.

ACKNOWLEDGEMENTS

We thank Steve Kissane (IBR, Birmingham, UK) and Anthony Jones (School of Biosciences, Birmingham, UK) for developing and manufacturing arrays. We thank Tim Williams (School of Biosciences, Birmingham, UK) for help with the software application Genespring. We also thank Ana Conesa (IVIA, Spain) and David Lowe (Aston University, Birmingham, UK) for very useful comments on the manuscript. D.S. is a recipient of a fellowship sponsored by The Darwin Trust. Funding to pay the Open Access publication charges for this article was provided by The Darwin Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

2. Chin,K.V. and Kong,A.N. (2002) Application of DNA microarrays in pharmacogenomics and toxicogenomics. *Pharm. Res.*, **19**, 1773–1778.
3. Butte,A. (2002) The use and analysis of microarray data. *Nature Rev. Drug. Discov.*, **1**, 951–960.
4. Heller,M.J. (2002) DNA microarray technology: devices, systems, and applications. *Annu. Rev. Biomed. Eng.*, **4**, 129–153.
5. Shirota,Y., Kaneko,S., Honda,M., Kawai,H.F. and Kobayashi,K. (2001) Identification of differentially expressed genes in hepatocellular carcinoma with cDNA microarrays. *Hepatology*, **33**, 832–840.
6. Okabe,H., Satoh,S., Kato,T., Kitahara,O., Yanagawa,R., Yamaoka,Y., Tsunoda,T., Furukawa,Y. and Nakamura,Y. (2001) Genome-wide analysis of gene expression in human hepatocellular carcinomas. *Cancer Res.*, **61**, 2129–2137.
7. Parsonage,G., Falciani,F., Burman,A., Filer,A., Ross,E., Bofill,M., Martin,S., Salmon,M. and Buckley,C.D. (2003) Global gene expression profiles in fibroblasts from synovial, skin and lymphoid tissue reveals distinct cytokine and chemokine expression patterns. *Thromb. Haemost.*, **90**, 688–697.
8. Iyer,V.R., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C., Trent,J.M., Staudt,L.M., Hudson,J.,Jr, Boguski,M.S. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.
9. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
10. Vlachonasios,K.E., Thomashow,M.F. and Triezenberg,S.J. (2003) Disruption mutations of ADA2b and GCN5 transcriptional adaptor genes dramatically affect *Arabidopsis* growth, development, and gene expression. *Plant Cell*, **15**, 626–638.
11. Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.
12. Stekel,D.J. (2003) *Microarray Bioinformatics*. Cambridge University Press, Cambridge, UK.
13. Jain,N., Thatte,J., Braciale,T., Ley,K., O'Connell,M. and Lee,J.K. (2003) Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, **19**, 1945–1951.
14. Dresen,I.M., Husing,J., Kruse,E., Boes,T. and Jockel,K.H. (2003) Software packages for quantitative microarray-based gene expression analysis. *Curr. Pharm. Biotechnol.* **4**, 417–437.
15. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H. and He,Y.D. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
16. Rocke,D.M. and Durbin,B. (2001) A model for measurement errors for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.
17. Love,B., Rank,D.R., Penn,S.G., Jenkins,D.A. and Thomas,R.S. (2002) A conditional density error model for the statistical analysis of microarray data. *Bioinformatics*, **18**, 1064–1072.
18. Strimmer,K. (2003) Modeling gene expression measurement error: a quasi-likelihood approach. *BMC Bioinformatics*, **4**, 10.
19. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
20. Donnenberg,M.S. and Whittam,T.S. (2001) Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic *Escherichia coli*. *J. Clin. Invest.*, **107**, 539–548.
21. Frankel,G., Phillips,A.D., Rosenshine,I., Dougan,G., Kaper,J.B. and Knutton,S. (1998) Enteropathogenic and enterohaemorrhagic *Escherichia coli*: more subversive elements. *Mol. Microbiol.*, **30**, 911–921.
22. Kenny,B. and Jepson,M. (2000) Targeting of an enteropathogenic *Escherichia coli* (EPEC) effector protein to host mitochondria. *Cell. Microbiol.*, **2**, 579–590.
23. Kenny,B., Ellis,S., Leard,A.D., Warawa,J., Mellor,H. and Jepson,M.A. (2002) Co-ordinate regulation of distinct host cell signalling pathways by multifunctional enteropathogenic *Escherichia coli* effector molecules. *Mol. Microbiol.*, **44**, 1095–1107.
24. McNamara,B.P., Koutsouris,A., O'Connell,C.B., Nougayrede,J.P., Donnenberg,M.S. and Hecht,G. (2001) Translocated EspF protein from enteropathogenic *Escherichia coli* disrupts host intestinal barrier function. *J. Clin. Invest.*, **107**, 621–629.
25. Crane,J.K., McNamara,B.P. and Donnenberg,M.S. (2001) Role of EspF in host cell death induced by enteropathogenic *Escherichia coli*. *Cell Microbiol.*, **3**, 197–211.
26. Sears,C.L. and Kaper,J.B. (1996) Enteric bacterial toxins: mechanisms of action and linkage to intestinal secretion. *Microbiol. Rev.*, **60**, 167–215.
27. Skinner,L.M. and Jackson,M.P. (1998) Inhibition of prokaryotic translation by the Shiga toxin enzymatic subunit. *Microb. Pathog.*, **24**, 117–122.
28. Malstrom,C. and James,S. (1998) Inhibition of murine splenic and mucosal lymphocyte function by enteric bacterial products. *Infect. Immun.*, **66**, 3120–3217.
29. Klapproth,J.M., Donnenberg,M.S., Abraham,J.M. and James,S.P. (1996) Products of enteropathogenic *E.coli* inhibit lymphokine production by gastrointestinal lymphocytes. *Am. J. Physiol.*, **271**, G841–G848.
30. Klapproth,J.M., Scaletsky,I.C., McNamara,B.P., Lai,L.C., Malstrom,C., James,S.P. and Donnenberg,M.S. (2000) A large toxin from pathogenic *Escherichia coli* strains that inhibits lymphocyte activation. *Infect. Immun.*, **68**, 2148–2155.
31. Knutton,S., Baldwin,T., Williams,P.H. and McNeish,A.S. (1989) Actin accumulation at sites of bacterial adhesion to tissue culture cells: basis of a new diagnostic test for enteropathogenic and enterohemorrhagic *Escherichia coli*. *Infect. Immun.*, **57**, 1290–1298.
32. Hegde,P., Qi,R., Abernathy,K., Gay,C., Dharap,S., Gaspard,R., Hughes,J.E., Snesrud,E., Lee,N. and Quackenbush,J. (2000) A concise guide to cDNA microarray analysis. *Biotechniques*, **29**, 548–550, 552–554, 556.
33. Yang,Y.K., Dudoit,S., Liu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T. (2002) Normalization for cDNA microarray data: a robust and composite method addressing single and multiple slide systemic variation. *Nucleic Acids Res.*, **30**, e15.
34. Quackenbush,J. (2002) Microarray data normalization and transformation. *Nature Genet.*, **32**(Suppl. 2), 496–501.
35. Efron,B. and Tibshirani,R.J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL, pp. 202–236.
36. Herrero,J., Vaquerizas,J.M., Al-Shahrour,F., Conde,L., Mateos,A., Santoyo,J., Díaz-Uriarte,R. and Dopazo,J. (2004) New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res.*, **32**, W485–W491.
37. Cleary,J., Lai,L.C., Shaw,R.K., Straatman-Iwanowska,A., Donnenberg,M.S., Frankel,G. and Knutton,S. (2004) Enteropathogenic *Escherichia coli* (EPEC) adhesion to intestinal epithelial cells: role of bundle-forming pili (BFP), EspA filaments and intimin. *Microbiology*, **150**(Pt 3), 527–538.
38. Baltathakis,I., Alcantara,O. and Boldt,D.H. (2001) Expression of different NF-kappaB pathway genes in dendritic cells (DCs) or macrophages assessed by gene expression profiling. *J. Cell. Biochem.*, **83**, 281–290.
39. Dahan,S., Busuttil,V., Imbert,V., Peyron,J.F., Rampal,P. and Czerucka,D. (2002) Enterohemorrhagic *Escherichia coli* infection induces interleukin-8 production via activation of mitogen-activated protein kinases and the transcription factors NF-kappaB and AP-1 in T84 cells. *Infect. Immun.*, **70**, 2304–2310.
40. Savkovic,S.D., Koutsouris,A. and Hecht,G. (1997) Activation of NF-kappaB in intestinal epithelial cells by enteropathogenic *Escherichia coli*. *Am. J. Physiol.*, **273**, C1160–C1167.
41. Wang,J., Jing,Z., Zhang,L., Zhou,G., Braun,J., Yao,Y. and Wang,Z.Z. (2003) Regulation of acetylcholine receptor clustering by the tumor suppressor APC. *Nature Neurosci.*, **6**, 1017–1018.
42. Khan,A.A., Bose,C., Yam,L.S., Soloski,M.J. and Rupp,F. (2001) Physiological regulation of the immunological synapse by agrin. *Science*, **292**, 1681–1686.
43. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate under dependency. *Ann. Statist.*, **29**, 1165–1188.
44. Reiner,A., Yekutieli,D. and Benjamini,Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.
45. La Ferla,K., Seegert,D. and Schreiber,S. (2004) Activation of NF-kappaB in intestinal epithelial cells by *E.coli* strains isolated from the colonic mucosa of IBD patients. *Int. J. Colorectal Dis.*, **19**, 334–342.
46. Innocenti,M., Thoreson,A.C., Ferrero,R.L., Stromberg,E., Bolin,I., Eriksson,L., Svennerholm,A.M. and Quiding-Jarbrink (2002) *Helicobacter pylori*-induced activation of human endothelial cells. *Infect. Immun.*, **70**, 4581–4590.

47. Hauf,N. and Chakraborty,T. (2003) Suppression of NF-kappa B activation and proinflammatory cytokine expression by Shiga toxin-producing *Escherichia coli*. *J. Immunol.*, **170**, 2074–2082.
48. Ceponis,P.J., McKay,D.M., Ching,J.C., Pereira,P. and Sherman,P.M. (2003) Enterohemorrhagic *Escherichia coli* O157:H7 disrupts Stat1-mediated gamma interferon signal transduction in epithelial cells. *Infect. Immun.*, **71**, 1396–1404.
49. DeVinney,R., Puente,J.L., Gauthier,A., Goosney,D. and Finlay,B.B. (2001) Enterohaemorrhagic and enteropathogenic *Escherichia coli* use a different Tir-based mechanism for pedestal formation. *Mol. Microbiol.*, **41**, 1445–1458.
50. McWhirter,J.R., Neuteboom,S.T., Wancewicz,E.V., Monia,B.P., Downing,J.R. and Murre,C. (1999) Oncogenic homeodomain transcription factor E2A-Pbx1 activates a novel WNT gene in pre-B acute lymphoblastoid leukemia. *Proc. Natl Acad. Sci. USA*, **96**, 11464–11469.